Improving Vision-and-Language Reasoning via Spatial Relations Modeling

Cheng Yang[1,*], Rui Xu[2,*], Ye Guo[3], Peixiang Huang[2], Yiru Chen[3], Wenkui Ding[3],
Zhongyuan Wang[3], Hong Zhou[1,†]
Zhejiang University[1], Peking University[2], Kuaishou Technology[3]
zijingyang@zju.edu.cn, {xurui,huangpx}@stu.pku.edu.cn
{guoye03,chenyiru,dingwenkui,wangzhongyuan}@kuaishou.com, zhouh@mail.bme.zju.edu.cn
*(equal contribution) †(corresponding author)

Abstract

Visual commonsense reasoning (VCR) is a challenging multi-modal task, which requires high-level cognition and commonsense reasoning ability about the real world. In recent years, large-scale pre-training approaches have been developed and promoted the state-of-the-art performance of VCR. However, the existing approaches almost employ the BERT-like objectives to learn multi-modal representations. These objectives motivated from the text-domain are insufficient for the excavation on the complex scenario of visual modality. Most importantly, the spatial distribution of the visual objects is basically neglected. To address the above issue, we propose to construct the spatial relation graph based on the given visual scenario. Further, we design two pre-training tasks named object position regression (OPR) and spatial relation classification (SRC) to learn to reconstruct the spatial relation graph respectively. Quantitative analysis suggests that the proposed method can guide the representations to maintain more spatial context and facilitate the attention on the essential visual regions for reasoning. We achieve the state-of-the-art results on VCR and two other vision-and-language reasoning tasks VQA, and NLVR^2.

1. Introduction

Vision-and-language reasoning is one of the most challenging tasks in multi-modal area, and the representative benchmarks incorporate Visual Commonsense Reasoning (VCR) [40], Visual Question Answering (VQA) [2] and Natural Language for Visual Reasoning (NLVR) [33]. Different from VQA and NLVR, VCR task requires to select the correct answer and provide corresponding explanation simultaneously given an image-question pair. Consequently, the comprehensive cognition-level scene understanding and cross-modal reasoning are essential for VCR.

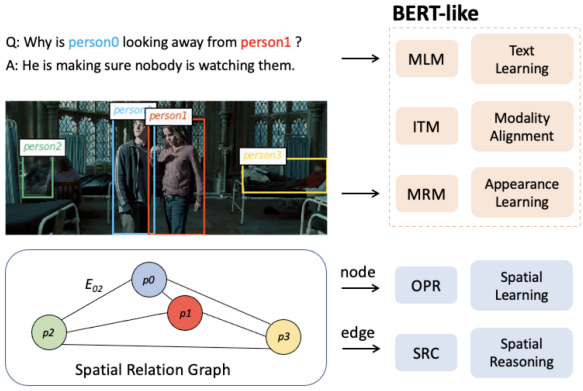

Figure 1. Existing pre-training approaches are insufficient for the excavation on visual modality. In this paper, we propose the spatial relation graph based on the visual modality and design two pre-training tasks named object position regression (OPR) and spatial relation classification (SRC) to promote the spatial context understanding and vision-and-language reasoning.

With the development of vision-and-language pre-training models in recent years [4, 21, 24, 39], state-of-the-art VCR algorithms basically follow the pretrain-and-finetune manners.

The existing vision-and-language pre-training approaches almost employ the BERT-like objectives to learn multi-modal representations , such as Masked Region Modeling (MRM) [24] similar to Masked Language Modeling (MLM) and Image-Text Matching (ITM) [4, 24, 32] similar to Next Sentence Prediction. Fig.1 illustrates the widely applied tasks for vision-and-language pre-training. Thereinto, MLM is to predict the masked text embeddings and ITM is to distinguish the matching of the image-text pair, lacking detailed excavation on the visual modality. The task of MRM, which is designed to classify the region feature extracted by the object detector, finitely concentrates on the individual semantic category and the appearance learning.

These BERT-like objectives motivated from text domain self-training are insufficient for the excavation on the complex scenario of visual modality. VCR task requires an in-depth understanding of the visual scenario and the commonsense reasoning beyond that. As Fig.1 shows, selecting the correct answer of the visualized case basically requires the spatial-aware capture on the "persons", which is less excavated in previous pre-training. Consequently, a more comprehensive understanding of the spatial context will significantly benefit the multi-modal reasoning for VCR.

To model the spatial context of the given scenario, we propose to construct the spatial relation graph directly with the coordinates of the visual object regions. Compared with previous methods [19, 36] introducing semantic-aware relations, our proposed method is free of external knowledge and training on the visual relationships. As Fig.1 shows, objects from the annotations and detection predictions constitute the graph nodes set. The value of each node is the corresponding region coordinates. The relations calculated with the coordinates of the object regions are represented as the graph edges. Beyond the BERT-like pre-training approaches, we propose to alternately learn on the constructed graph, promoting the spatial relations modeling on the multi-modal data.

Concretely, we design two novel pre-training tasks to recover the property of nodes and edges in the constructed spatial relation graph respectively. Existing vision-and-language pre-training methods [4, 10, 21, 39] almost adopt BUTD [1] to extract the object visual embeddings. Among them, the region positions are just finitely utilized as an auxiliary input to the visual embeddings. To maintain more spatial information in the multi-modal representations, we propose alternative pre-training with Object Position Regression (OPR) and Spatial Relation Classification (SRC). Taking the textual data and visual features as the context, OPR is to predict the masked positions pruning the input position information for each object. Beyond the individual spatial modeling, SRC aims to explicitly raise awareness of the spatial relations among the object. Noteworthy, the proposed alternative pre-training is significantly different with previous methods [19, 38] introducing visual relations, which followed by graph-based networks for visual representation learning. We are the first to regard the spatial relation graph as learning targets of the multi-modal pre-training, which can be easily applied on current universal and advanced transformer-based frameworks.

Experimental results demonstrate alternative training with OPR and SRC can achieve a significant performance boost compared with previous state-of-the-art methods for VCR. Quantitative analysis suggests the proposed pre-training tasks can guide the representations to capture more spatial information and improve the attention weights on more essential visual regions for reasoning. Additional experiments on VQA and NLVR^2 further prove the effectiveness of our method in vision-and-language reasoning field.

The contributions of our method are three-folds:

- To the best of our knowledge, we are the first to regard the spatial relation graph as learning targets, and promote spatial context understanding of the vision-and-language representations.
- We propose two novel pre-training tasks, named Object Position Regression and Spatial Relation Classification, which can be widely applied on universal transformer-based multi-modal frameworks without external knowledge.
- We achieve the state-of-the-art results among the models of comparable scale. Experiments are conducted on VCR (with a significant improvement compared with previous works) and two other vision-and-language reasoning tasks VQA, and NLVR^2.

2. Related Work

Representation Learning. In recent years, there are substantial interests in both vision [3, 7, 9, 11] and language [6, 8, 29] pre-training for representation learning. Most visual pre-training methods are based on the convolutional neural network architecture (CNN) such as VGG [31] and ResNet [12] trained on the ImageNet dataset [5]. The language pre-training methods are almost based on multi-layer transformer [35]. BERT introduces Masked Language Modeling (MLM) pre-training task that randomly masks the input words and predicts these masked words based on the contexts. MLM has been a standard schema for linguistic model representation learning.

Vision-and-Language Representation Learning. ViLBERT [24] and LXM-ERT [34] are the pioneering works in vision-and-language representation learning, where two parallel transformers are utilized to process visual features or language embeddings separately, and a third transformer is built on the top for multi-modal features fusion. Compared to the above architecture, recent work such as VisualBERT [20], VL-BERT [32], Unicoder-V [18] and UNITER [4] advocate a single-stream architecture, where two modalities are fused in the early stage. VinVL [42] improves the vision-and-language models by developing an improved object detection model to generate object-centric representations of images. SOHO [13] learns to extract comprehensive image features through a visual dictionary that facilitates cross-modal understanding. CATT [37] proposes causal attention to remove the ever-elusive confounding effect in the existing attention-based models. Moreover, other techniques like knowledge integration [39], con-

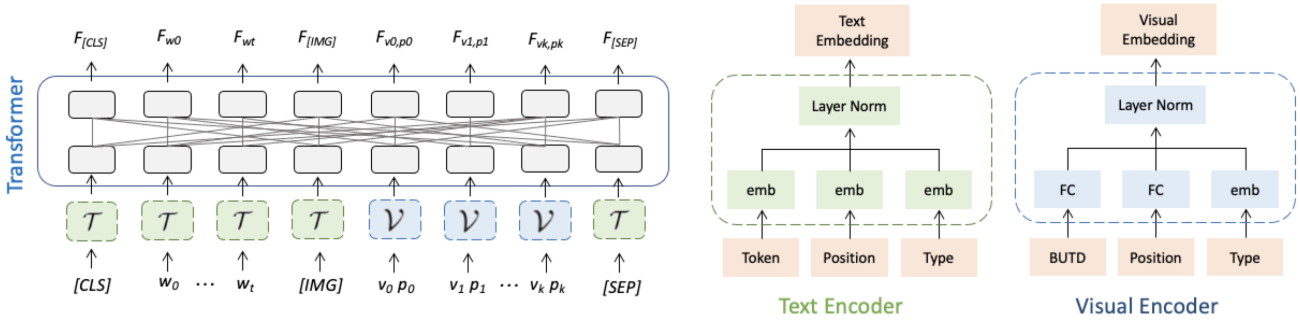

Figure 2. Overview of the model. We employ the multi-layer self-attention transformer to learn the multi-modal representation for both textual and visual data. The text encoder $\mathcal{T}$ aggregates the token, position, type of each word w_i into the input text embeddings. The visual encoder $\mathcal{V}$ aggregates the object feature extracted by BUTD [1], bounding box coordinates, type of each region v_i into the input visual embeddings.

trastive learning [21], adversarial training [10], supervision from text [14, 28], modality matching from large scale video dataset [41] are introduced to further improve the performance of the pre-trained models. The above models have brought leaping advances in vision-and-language downstream tasks such as VCR [40], visual captioning [26, 27, 43], visual dialog [25] and image-text retrieval [17].

Vision-and-Language Pre-training Tasks. Most of pre-training approaches directly employ BERT-like objectives to learn multi-modal representations, such as Masked Region Model (MRM) [24] similar to Masked Language Model (MLM) and Image-Text Matching (ITM) [4, 24, 32] similar to Next Sentence Prediction (NSP). The ordinary MLM pre-training neglects the semantic relationships among the textual data. ERNIE-ViL [39] improves the masking strategy by predicting the token types according to the textual scene graph, incorporating objects, attributes, and relationships. By increasing the corresponding masking probability, ERNIE-ViL achieves a better semantic alignment across the vision and language modality. The ordinary ITM pre-training randomly samples a negative image or text from the same training batch for each pair, leading to a coarse alignment between the textual and visual representations. To further facilitate the alignment, UNIMO [21] proposes text rewriting techniques to augment the original captions at word, phrase, and sentence levels. In this way, UNIMO obtains large volumes of positive and negative examples for each image-text pair. Furthermore, cross-modal contrastive learning (CMCL) is leveraged by UNIMO to align the textual and visual information into a unified semantic space.

Visual Relationship Enhanced Representation Learning. Previous methods [15, 19, 36, 38] introduce scene graphs to model the spatial or semantic relationships among the visual objects, almost followed by graph-based attention networks to enhance the visual features. However, we propose to construct the spatial relation graph without external semantic information, are the first to regard the constructed spatial graph as learning targets of multi-modal pre-training, promoting spatial context understanding of the vision-and-language representations. Our proposed spatial relations modeling can be easily applied on large-scale and universal transformer frameworks with designed masked strategies and loss functions. Experimental results demonstrate that our method gains significant improvement on several benchmarks.

3. Approach

In this section, we first introduce the architecture of our model. Then we illustrate the construction of spatial relation graph and the proposed pre-training tasks for spatial relations modeling. Finally, we describe the complete pre-training objectives and procedures with alternative learning.

3.1. Overview of the Pre-trained Model

The vision-and-language pre-trained model aims at learning the joint representations that integrate information of both visual and textual modalities. As shown in Fig.2, we employ multi-layer transformer [6] to learn the unified representations. For texts data $\mathbf{w}=\{w_i\}$, we adopt the token embedding initialized by RoBERTa [23], and special tokens incorporating $[CLS]$, $[IMG]$ and $[SEP]$ are added to the tokenized sequences. The texts are divided into different types for the questions or answers. With the text encoder $\mathcal{T}$ in Fig.2 shows, the input text embedding for each sub-word is generated by aggregating its original token embedding, sequence position embedding, and type embedding.

Similarly, the image is also converted to a sequence of visual embeddings. Consistent with [4, 21, 39], we use BUTD [1] to detect the foreground regions and extract the visual features correspondingly, denoted as $\mathbf{p}=\{p_i\}$ and $\mathbf{v}=\{v_i\}$ respectively. The position information for each object is encoded via a 5-dimensional vector p_i as

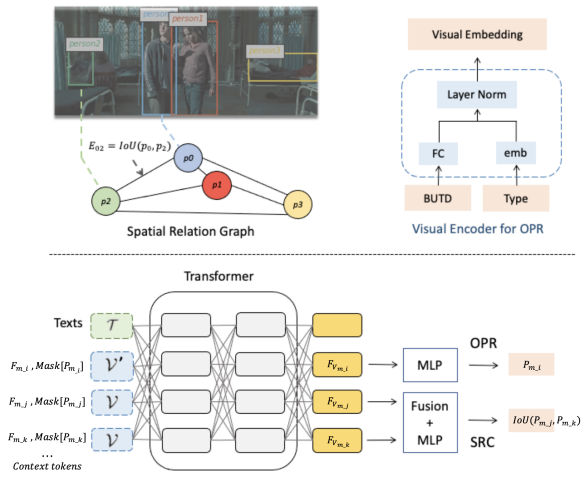

Figure 3. We propose to construct the spatial relation graph for the visual scenario. Based on the graph, we design two novel pre-training tasks named Object Position Regression (OPR) and Spatial Relation Classification (SRC) learning to recover the property of the nodes and edges respectively.

Equation (1) shows, where (x_1, y_1) and (x_2, y_2) denote the region top-left and bottom-right corner coordinates, while W and H denote the width and height of the image. The position vectors are projected to the same dimension as $\mathbf{v}$. With the visual encoder $\mathcal{V}$, the input visual embedding for each region is generated by aggregating its visual feature, position embedding, and the visual type embedding. For a text-image pair, its textual and visual tokens are concatenated as a sequence. Then the sequence input embeddings are feed into the multi-layer transformer to learn the final multi-modal representations: $\{F_{[CLS]}, F_{w0}, \ldots, F_{w_t}, F_{[IMG]}, F_{v_0,p_0}, \ldots, F_{v_k,p_k}, F_{[SEP]}\}$.

$$p_i = (\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \frac{(y_2 - y_1)(x_2 - x_1)}{WH}) \quad (1)$$

3.2. Spatial Relations Modeling

The existing vision-and-language pre-training tasks partially concentrate on the individual category and appearance learning. For a complex visual scenario, objects positions and the interactive spatial relations contain more information to be excavated for multi-modal reasoning.

Spatial Relation Graph. To model the spatial context of the scenario, we propose to construct the spatial relation graph for the given images. As Fig.3 shows, objects from annotations or detection predictions constitute the graph nodes set $\mathbf{p}$. The value of each node is the corresponding object spatial position vectors p_i in Equation (1). The relations calculated with the position vectors are represented as the spatial relation edges $\{E_{ij}\}$.

We investigate various spatial relation descriptions from the aspect of directions or overlap between the visual objects, as Table.6 shows. Eventually, we introduce the Intersection over Union (IoU) to quantify the spatial relations. We design two novel pre-training tasks named Object Position Regression (OPR) and Spatial Relation Classification (SRC) to learn to reconstruct the nodes and edges of the graph respectively, promoting the spatial context and comprehensive cross-model understanding.

Object Position Regression (OPR). As Fig.2 shows, in the conventional visual encoder $\mathcal{V}$, position features are finitely utilized as auxiliary information in the input embeddings. Guiding the representations to maintain more spatial information, OPR is designed to predict the position vector of each object pruning the original input with the textual and visual appearance clues as context.

Concretely, to guarantee the precision of the targeted position vector, only objects with detection confidence scores larger than 0.5 and ground truth objects are available to be masked. For OPR pre-training, we randomly mask the available position vectors with a probability 50%. As Fig.3 shows, the visual encoder for OPR is fed with only feature embedding and token type embedding, excluding bounding box position embedding, for the position-masked objects. With additional two layers MLP network, the transformer output F_{v_m} is projected to predict the masked position vector p_m. We denote the object features as $\mathbf{v}$, the corresponding position vectors as $\mathbf{p}$, and the input words as $\mathbf{w}$. Each image-text pair $(\mathbf{v}, \mathbf{p}, \mathbf{w})$ is sampled from the whole training set D. The model parameters set is denoted as θ and the predicted position vector can be denoted as $P_\theta(p_m|\mathbf{v}, \mathbf{p}_{\setminus m}, \mathbf{w})$. The task directly predicts the 5-dimensional vector in Equation (1). The loss function of OPR can be summarized as follows.

$$\mathcal{L}_{OPR}(\theta) = \mathbb{E}_{(\mathbf{v},\mathbf{p},\mathbf{w})\in D} ||P_\theta(p_m|\mathbf{v}, \mathbf{p}_{\setminus m}, \mathbf{w}) - p_m||^2 \quad (2)$$

Spatial Relation Classification (SRC). As stated above, OPR guides the pre-trained model to maintain more individual spatial information of each object, i.e nodes of the spatial relation graph in Fig.3. For further modeling the spatial context, we propose to learn to reconstruct the edges of spatial relation graph, namely spatial relation classification (SRC).

To model the relations, we detailedly investigate the effect on the performance with different spatial metrics and modeling approaches (Table 6). Eventually, we introduce the IoU from the overlapping aspect to model the relationships, i.e E_{ij}=$IoU(p_i, p_j)$. Since IoU regression is an excessively tough task for the continuous variation, we model the prediction of the IoU as a classification problem. Concretely, we divide the IoU $\{E_{ij}\}$ to 10 classes with a uniform interval of 0.1, and the constant target is

transferred to a category target E'_{ij}. For SRC pre-training, a pair of visual objects (p_i, p_j) are sampled. The corresponding transformer output $(F_{v_i,p_i}, F_{v_j,p_j})$ are fused and projected by a two-layer MLP network to predict the relationship label E'_{ij}. The predicted spatial relation is denotes as $P_\theta(E'_{ij}|\mathbf{v}, \mathbf{p}, \mathbf{w})$ and softmax-based cross-entropy loss is adopted. The loss function of SRC pre-training can be summarized as follows.

$$\mathcal{L}_{SRC}(\theta) = -\mathbb{E}_{(\mathbf{v},\mathbf{p},\mathbf{w})\in D} log P_\theta(E'_{ij}|\mathbf{v}, \mathbf{p}, \mathbf{w}) \quad (3)$$

3.3. Alternative pre-training

We propose to alternatively train the multi-modal model with the proposed OPR and SRC from the aspect of spatial context, combining Masked Language Modeling (MLM) and Masked Region Classification (MRC) from the aspect of semantic meanings. By the means of alternative pre-training, the model can capture more comprehensive representations for the visual and textual modality. Next, we will briefly introduce the implementation details and objectives of the adopted MLM and MRC.

For Masked Language Modeling, we randomly mask the input textual token embeddings with a probability of 15%, and replace the masked $\mathbf{w}_m$ with a special token [MASK]. The model is trained to predict the masked tokens based on the surrounding context. Denote the prediction is $P_\theta(w_m|\mathbf{v}, \mathbf{p}, \mathbf{w}_{\setminus m})$, and the softmax-based cross-entropy loss function can be summarized as follows.

$$\mathcal{L}_{MLM}(\theta) = -\mathbb{E}_{(\mathbf{v},\mathbf{p},\mathbf{w})\in D} log P_\theta(w_m|\mathbf{v}, \mathbf{p}, \mathbf{w}_{\setminus m}) \quad (4)$$

For Masked Region Classification, we randomly sample image regions v_m and mask out their visual features with a probability of 15%. The model is trained to predict the object category of each masked region based on the context. Additional fully-connected layers are introduced to project the prediction to the categories probability distribution $P_\theta(v_m|\mathbf{v}_{\setminus m}, \mathbf{p}, \mathbf{w})$. We adopt the KL-divergence loss function to minimize the difference between the prediction and the object detection model labeled distribution $\hat{v}_m$. The loss function of MRC can be summarized as follows.

$$\mathcal{L}_{MRC}(\theta) = \mathbb{E}_{(\mathbf{v},\mathbf{p},\mathbf{w})\in D} KL(P_\theta(v_m|\mathbf{v}_{\setminus m}, \mathbf{p}, \mathbf{w}), \hat{v}_m) \quad (5)$$

4. Experiments

4.1. Dataset

Visual Commonsense Reasoning (VCR) dataset [40] contains 100K images and 264K related questions, which are divided into the *train*, *val*, and *test* split at a ratio of 8:1:1. The VCR task incorporates two sub-tasks: visual question answering (*Q*→*A*) and answer justification (*QA*→*R*), which are both multiple-choice problems. The holistic setting (*Q*→*AR*) requires both the chosen answer and the chosen rationale to be correct. In the visual question answering (*Q*→*A*) task, we concatenate the question and each candidate answer for the language modality. We take dot product of the final transformer outputs $F_{[CLS]}$ and $F_{[IMG]}$ to predict the matching score with an additional FC layer. For the answer justification (*QA*→*R*) task, we concatenate the question, the answer, and each candidate rationale as the input of the textual data.

Table 1. Results of the VCR task compared with the previous state-of-the-art models.

Models	VCR *val*					
	base			large		
	Q→*A*	*QA*→*R*	*Q*→*AR*	*Q*→*A*	*QA*→*R*	*Q*→*AR*
ViLBERT [24]	72.4	74.5	54.0	-	-	-
VisualBERT [20]	70.8	73.2	52.2	-	-	-
SGEITL [36]	-	-	-	74.9	77.2	57.8
VL-BERT [32]	73.8	74.4	55.2	75.5	77.9	58.9
UNITER [4]	74.6	77.0	57.8	77.2	80.5	62.6
VILLA [10]	75.5	78.8	59.8	78.5	82.6	65.2
ERNIE-ViL [39]	76.4	79.7	61.2	79.0	83.7	66.4
Ours	**78.8**	**83.1**	**65.8**	**83.0**	**87.9**	**73.4**

4.2. Implementation Details

For fair comparison with previous methods, the experiments are conducted on two model sizes: base with 12 layers of transformer block and large with 24 layers of transformer block. We initialize the pre-trained model with UNIMO [21] for the main results, and conduct another stage of pre-training on the training split of VCR. The multi-task mix ratios for OPR, SRC, MLM and MRC are 1:1:10:1. The total number of pre-training steps is 50,000. After the alternative pre-training, we fine-tune the model over 10,000 steps with a batch size of 64 and adopt Adamw optimizer with an initial learning rate of 6e-4.

For all experiments, we use AdamW optimizer with weight decay of 10^{-2}. The learning rate is warmed up for 10% of the total training steps and is decayed linearly to zero for the rest of the training. The maximum sequence length of text tokens and visual regions are set as 514 and 100, respectively. For the text tokens, we adopt Byte-Pair Encoding (BPE) to tokenize the sentence similar to RoBERTa [23]. For the visual regions, we adopt BUTD [1] pre-trained on the Visual Genome [16] to detect the object regions and extract the visual features (pooled ROI features) correspondingly, which is consistent with previous methods [4, 21, 39]. Specifically, regions with class detection probability exceed a confidence threshold of 0.2 are selected. For the masking strategies, we randomly mask 15% of tokens in MLM, 15% of object region features in

MRC, and 50% of region position vectors in OPR.

4.3. Main Results

We compare our method against the previous state-of-the-art models and the results are illustrated in Table 1. We obtain a 4.6% and 7.0% improvement for *Q→AR* under the base and large setting compared with previous public state-of-the-art results achieved by ERNIE-ViL [39]. Meantime, our result significantly outperforms SGEITL [36], which introduce external scene graphs to enhance the representation learning.

Table 3 reports the *test* split evaluation results of VCR. Under the large setting, we achieve the state-of-the-art results with a single model, a 2.4% improvement even compared with the UNIMO [21] ensembling 7 models. VCR is a challenging task, which requires cross-modal commonsense reasoning and understanding on the complex scenario which is implicitly encoded in the image. Experimental results suggest we achieve new state-of-the-art results across the three benchmarks for VCR. The results strongly demonstrate the effectiveness of our method.

4.4. Ablation Study

To further analyze the effectiveness of OPR and SRC, we conduct detailed ablation studies and investigate the influence on VCR at different settings.

What type of pre-training task is more effective? We conduct the ablation study with different pre-training tasks on UNITER-base [4]. Previous methods [4, 10, 39] almost adopt MLM, MRFR and MRC for the pre-training on VCR. Thereinto, MRFR is a pre-training task similar to MRC, which directly regresses the object appearance features rather than predict the semantic categories. As Table 4 shows, the contribution of MRFR pre-training is almost useless for VCR, which suggests local appearance learning for the visual objects is already sufficient for the multi-modal pre-training.

However, the performance is significantly better if MRFR is replaced with either SRC or OPR. As Table.4 shows, when we conduct SRC and OPR simultaneously, a 1.0% improvement on QA→R and a 1.4% improvement on Q→AR can be obtained. Experimental results suggest spatial relations modeling is a valuable and vital complement for current multi-modal representation learning on visual modality. For further verification on the general effectiveness of OPR and SRC, we apply the proposed method on UNITER-large and VILLA-large [10]. As Table 5 shows, both UNITER and VILLA can obtain an improvement with OPR and SRC alternative pre-training. The ablation study on various pre-trained models convincingly demonstrates pre-training from the spatial perspective is necessary and effective for the multi-modal reasoning tasks.

How to model the spatial relations? We investigate different metrics for the spatial relations and various modeling approaches for SRC, as Table 6 shows. For direction prediction, we conduct left/right and upside/below classification between the visual object centers. For overlapping prediction, SRC learns to classify whether the sampled object regions are overlapped. Experimental results suggest overlapping prediction is a better metric than direction prediction. The supposed reason is the overlapping can reflect more relevance of the objects.

Further, we conduct more detailed investigation on the overlapping prediction with IoU and GIoU [30]. Experimental results suggest fine-grained classification on IoU can obtain a superior performance than binary classification on overlapping. In contrast, IoU regression is a excessively tough task, even harmful to the final performance. The performance of GIoU classification is slightly weaker than the trivial IoU metric, eventually we select IoU classification as the pre-training objective for SRC.

4.5. Experiments on Other Dataset

To prove the generalization of our method, we conduct experiments on other vision-and-language reasoning tasks, Visual Question Answering (VQA) [2] and Natural Language for Visual Reasoning (NLVR) [33]. Results compared with the state-of-the-art models are summarized in Table.2.

Visual Question Answering (VQA). VQA2.0 contains 204K images and 1.1M related questions, which are divided into the *train*, *val*, and *test* split at a ratio of 2:1:2. The VQA task requires answering natural language questions according to the given images. We treat VQA as a multi-label classification task assigning a soft target score to each answer based on its relevancy to the 10 human answer responses. We take dot product of the outputs $F_{[CLS]}$ and $F_{[IMG]}$ and map the representations into 3,129 possible answers with an additional two FC layers. We adopt the same pre-training schedule as VCR. Fine-tuning on VQA is performed over 5K steps with a batch size of 256 and we adopt the Adamw optimizer with an initial learning rate of 5e-4.

As Table.2 shows, we achieve new state-of-the-art results both on base and large setting. Concretely, we get a 0.2% performance boost under the base model setting evaluated on VQA *test-dev* and *test-std*. For the large model setting, the margin reaches up to 0.4% compared with UNIMO [21]. The results suggest the spatial relations modeling is also effective for question-answer tasks.

Natural Language for Visual Reasoning. NLVR^2 contains 107K examples of human-written English sentences, which are divided into the train, *dev*, *testP* and *testU* at a ratio of 12:1:1:1. The task is to determine whether a natural language caption is corresponding with a series of photographs. We take dot product of the final transformer outputs $F_{[CLS]}$ and $F_{[IMG]}$ to predict the matching score for

Table 2. Results of VQA and NLVR^2 compared with previous state-of-the-art models.

Models	VQA				NLVR^2			
	base		large		base		large	
	test-dev	*test-std*	*test-dev*	*test-std*	*dev*	*testP*	*dev*	*testP*
ViLBERT [24]	70.6	70.9	-	-	-	-	-	-
VisualBERT [20]	70.8	71.0	-	-	67.4	67.0	-	-
LXMERT [34]	72.4	72.5	-	-	74.9	74.5	-	-
LXMERT+CATT [37]	73.5	73.7	-	-	77.2	77.2	-	-
VL-BERT [32]	71.2	-	71.8	72.2	-	-	-	-
UNITER [4]	72.7	72.9	73.8	74.0	77.2	77.9	79.1	80.0
SOHO [13]	73.3	73.5	-	-	76.4	77.3	-	-
Oscar [22]	73.2	73.4	73.6	73.8	78.1	78.4	79.1	80.4
VILLA [10]	73.6	73.7	74.7	74.9	78.4	79.3	79.8	81.5
ERNIE-ViL [39]	73.2	73.4	75.0	75.1	-	-	-	-
UNIMO [21]	73.8	74.0	75.1	75.3	-	-	-	-
Ours	**74.0**	**74.2**	**75.5**	**75.6**	**79.4**	**80.1**	**80.6**	**82.2**

Table 3. The comparison of VCR *test* set evaluation under the large setting. $*$: the result is achieved by ensembling 7 models; partially from the VCR Leaderboard [40].

Models	VCR *test*		
	Q→*A*	*QA*→*R*	*Q*→*AR*
VL-BERT [32]	75.8	78.4	59.7
SGEITL [36]	76.0	78.0	59.6
UNITER [4]	77.3	80.8	62.8
VILLA [10]	78.9	82.8	65.7
ERNIE-ViL [39]	79.2	83.5	66.3
UNIMO* [21]	82.3	86.5	71.4
Ours	**83.2**	**88.1**	**73.8**

Table 4. The results of VCR *val* with different pre-training tasks on UNITER-base.

Alternative pre-training Tasks	*Q*→*A*	*QA*→*R*	*Q*→*AR*
MLM+MRC	74.3	76.9	57.3
MLM+MRC+MRFR	74.3	76.9	57.4
MLM+MRC+OPR	74.6	77.2	58.0
MLM+MRC+SRC	74.8	77.1	58.1
MLM+MRC+SRC+OPR	**75.2**	**77.9**	**58.7**

each image-text pair with an additional FC layer. For fine-tuning, we train the models with 5K steps totally and a batch size of 32. The Adamw optimizer with an initial learning rate 2e-5 is adopted.

As Table 2 shows, we achieve a 1.0% and 0.8% improvement on NLVR^2 *dev* under the base and large setting respectively. For NLVR^2 *testP*, the improvement is 0.8% and 0.7% correspondingly. Validation on NLVR^2 also support the conclusion stated above.

Table 5. The results of VCR *val* with OPR and SRC on UNTER and VILLA-large.

Models	*Q*→*A*	*QA*→*R*	*Q*→*AR*
UNITER	77.1	80.3	62.1
UNITER+OPR+SRC	**78.3**	**81.6**	**64.2**
VILLA	78.1	82.0	64.4
VILLA+OPR+SRC	**78.4**	**82.3**	**64.8**

Table 6. The results of VCR *val* with different SRC modeling methods.

SRC Modeling	*Q*→*A*	*QA*→*R*	*Q*→*AR*
Direction Prediction	81.3	87.1	71.0
Overlapping Prediction	81.5	87.0	71.2
IoU Regression	81.2	86.9	70.8
IoU Classification	**81.7**	**87.1**	**71.3**
GIoU Classification	81.3	**87.1**	71.1

Table 7. Average correlation coefficient between the position embeddings and the transformer input visual / output $F_{[CLS]}$ representations.

Alternative pre-training	Input Corr.	Output Corr.
MLM+MRC	0.14	-0.0060
MLM+MRC+SRC	**0.19**	0.0009
MLM+MRC+OPR	0.16	0.0020
MLM+MRC+OPR+SRC	0.19	**0.0041**

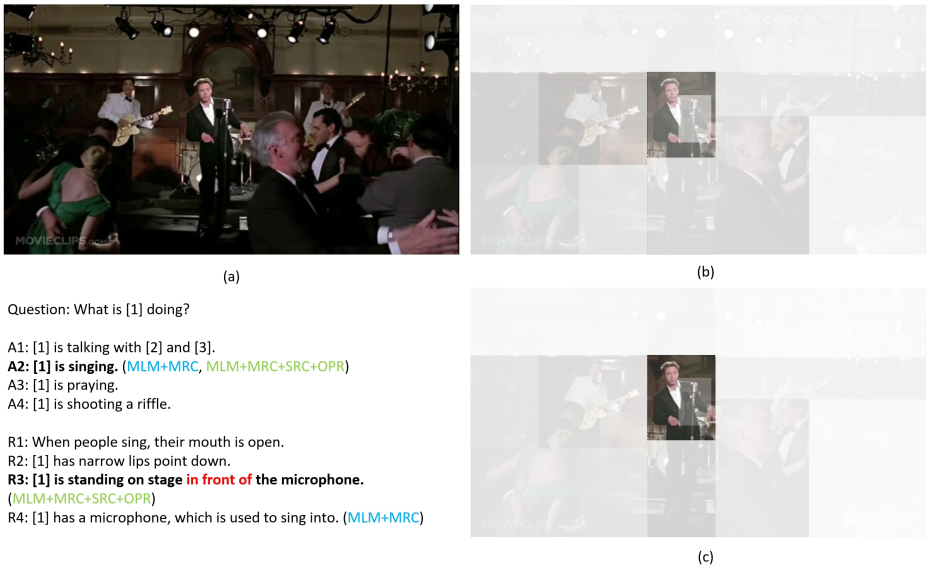

Figure 4. Case study on the text-to-object attentions (darker represents larger). (b) is the attention visualization of baseline, (c) is the attention with OPR and SRC pre-training. The correct answer and rationale are marked in bold. The answers picked by the models are indicated in parenthesis. pre-training with OPR and SRC, which memorizing more spatial relations clues in the visual scenario, can improve the attention weights on the more essential visual regions, thus benefits the multi-modal reasoning.

4.6. Discussion

In this section, we discuss the impact of spatial-aware modeling on the multi-modal representation learning and reasoning. By quantitative analysis on the features and attention weights, the motivation can be further demonstrated.

How spatial modeling impacts the representation learning? To explore how the proposed spatial perspective modeling impacts on the multi-modal representations, we conduct feature correlation analysis on 100 random VCR samples as Table 7 shows. Thereinto, the "Input Corr" denotes the average correlation coefficient between the input position embedding and the input visual representation $\mathcal{V}(v_k, p_k)$ for each object. The "Output Corr" denotes the average correlation coefficient between the input position embeddings and the output multi-modal representation $F_{[CLS]}$.

Quantitative results suggest the correlation coefficient stated above is boosted with OPR and SRC alternative pre-training. It can be inferred that the proposed spatial relations modeling can facilitate the maintaining and memorization of the spatial context before and after the transformer layers by alternative pre-training. The spatial information eventually can be exploited in the fine-tuning stage, thus benefits the multi-modal reasoning downstream tasks.

Attention weights analysis. Fig.4 provides an example of the learned text-to-object attentions. We can see that the baseline model selects the right answer, but the wrong rationale, which can be corrected with the proposed OPR and SRC pre-training. In this case, the spatial relation between the person[1] and the microphone is essential for the reasoning. The attention visualization suggests that pre-training with OPR and SRC, which memorizing more spatial relations clues, can improve the attention weights on the more relevant visual regions for the multi-modal reasoning.

5. Conclusion

Previous vision-and-language pre-training approaches motivated by text domain are insufficient on the visual modality excavation for reasoning. To address the above issue, we propose to construct the spatial relation graph based on the given visual scenario. Further, we design two pre-training tasks named object position regression (OPR) and spatial relation classification (SRC) to learn to reconstruct the graph respectively. By alternative pre-training with OPR and SRC, we achieve state-of-the-art results on three visual-and-language reasoning downstream tasks VCR, VQA, and NLVR^2. In particular, even though the VCR task is considered to be a very difficult multi-modality reasoning task, our method improves the performance of previous works by over 2.4%, which is a significant margin. Additionally, we also conduct detailed ablative experiments to demonstrate the effectiveness of our proposed pre-training tasks. Quantitative analysis suggests the spatial relations modeling can guide the model to maintain more spatial context and facilitate the attention on essential regions, thus benefits the challenging multi-modal reasoning. **This work is supported by the National Key Research and Development Program of China (2022YFC3602601).**

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2, 3, 5

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 6

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 1, 2, 3, 5, 6, 7

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. 2, 3

[7] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2

[8] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*, 2019. 2

[9] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1734–1747, 2015. 2

[10] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning, 2020. 2, 3, 5, 6, 7

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[13] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021. 2, 7

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3

[15] Yash Kant, Dhruv Batra, Peter Anderson, Alexander Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. Spatially aware multimodal transformers for textvqa. In *European Conference on Computer Vision*, pages 715–732. Springer, 2020. 3

[16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016. 5

[17] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 3

[18] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020. 2

[19] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10313–10322, 2019. 2, 3

[20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2, 5, 7

[21] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607, 2021. 1, 2, 3, 5, 6, 7

[22] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 7

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3, 5

[24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: pretraining task-agnostic visiolinguistic representations

for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23, 2019. 1, 2, 3, 5, 7

[25] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *European Conference on Computer Vision*, pages 336–352. Springer, 2020. 3

[26] Wenkang Qin, Rui Xu, Peixiang Huang, Xiaomin Wu, Heyu Zhang, and Lin Luo. What a whole slide image can tell? subtype-guided masked transformer for pathological image captioning. *arXiv preprint arXiv:2310.20607*, 2023. 3

[27] Wenkang Qin, Rui Xu, Shan Jiang, Tingting Jiang, and Lin Luo. Pathtr: Context-aware memory transformer for tumor localization in gigapixel pathology images. In *Proceedings of the Asian Conference on Computer Vision*, pages 3603–3619, 2022. 3

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[30] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 6

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[32] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019. 1, 2, 3, 5, 7

[33] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 1, 6

[34] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2, 7

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[36] Zhecan Wang, Haoxuan You, Liunian Harold Li, Alireza Zareian, Suji Park, Yiqing Liang, Kai-Wei Chang, and Shih-Fu Chang. Sgeitl: Scene graph enhanced image-text learning for visual commonsense reasoning. *arXiv preprint arXiv:2112.08587*, 2021. 2, 3, 5, 6, 7

[37] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9847–9857, 2021. 2, 7

[38] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018. 2, 3

[39] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216, 2021. 1, 2, 3, 5, 6, 7

[40] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019. 1, 3, 5, 7

[41] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae-Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Neural Information Processing Systems,Neural Information Processing Systems*, Dec 2021. 3

[42] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2

[43] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020. 3